PCA Example

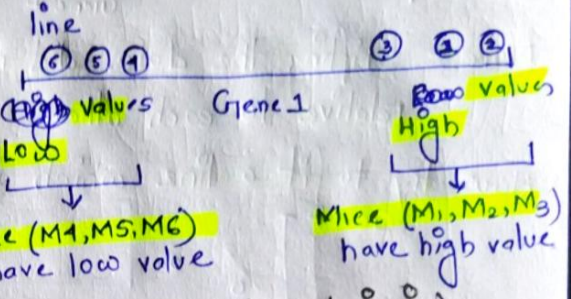
→ Suppose we have a dataset, ① Mice as individual sample = Sample 1, 2, 3.
② Gene as variable that we measure for each sample = Variable 1, 2, 3.

Suppose if we take one Gene, (One dimension)

	Mouse 1	M2	M3	M4	M5	M6
Gene 1	10	11	8	3	2	1

So from the number line we can say, it shows Mice 1, 2 and 3 are more similar to each other than they are to Mice 4, 5 and 6.

If we have one gene we can plot the data on a number line

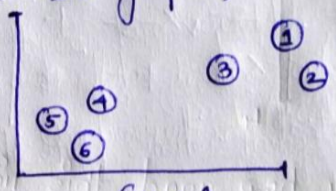


Suppose if we two Gene i.e. Gene 1 and Gene 2. (two dimension)

	M1	M2	M3	M4	M5	M6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

Mice (M1, M2, M3) cluster on right side.
Mice (M4, M5, M6) cluster on left side.

plot 2D graph,



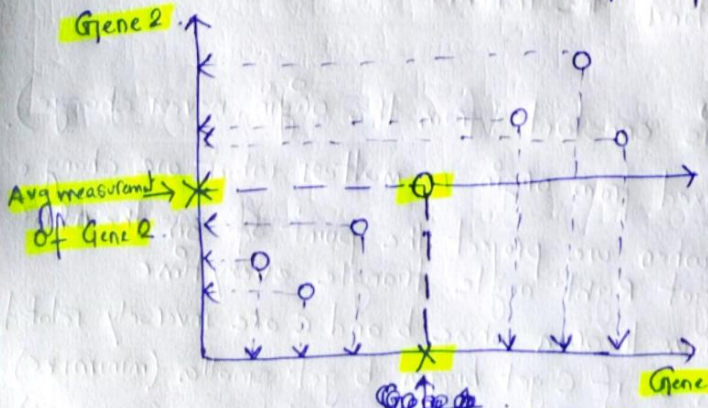
Suppose if we take 3 Gene i.e, Gene 1, Gene 2 and Gene 3. Use 3D graph (3 Dimensions) COMPLEX

So PCA can take many dimension (4 or more dimension) and make 2D plot.

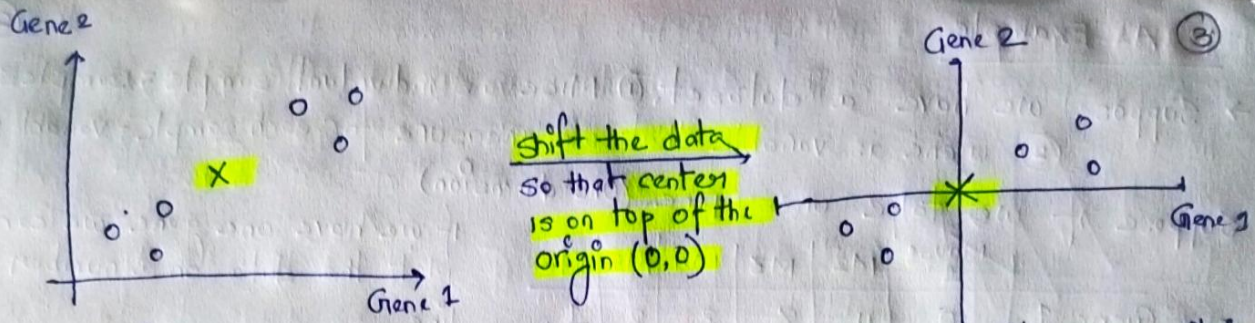
So how PCA works (step by step) →

- Consider an example of two gene,

	Mouse 1	M2	M3	M4	M5	M6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

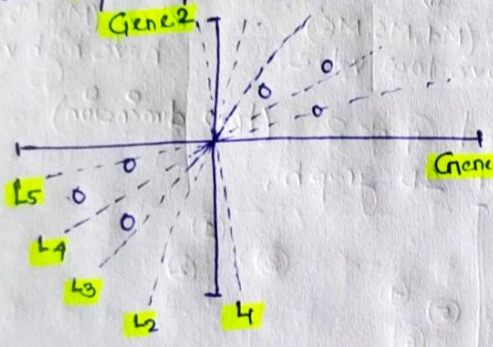


With average value, we can calculate the center of the data.



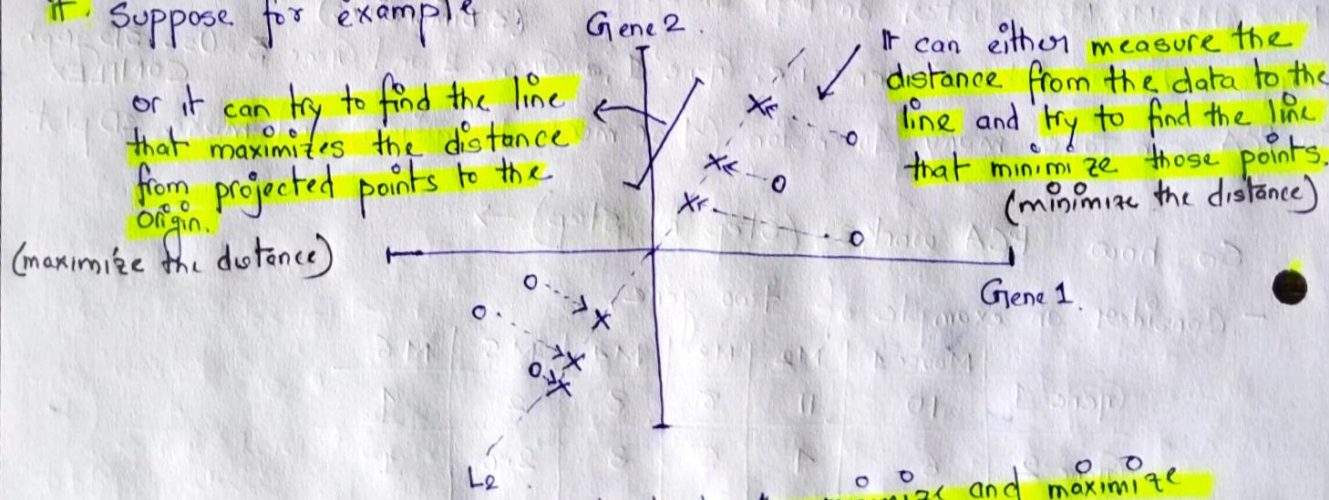
Note - shifting the data did not change how the data points are positioned relative to each other.

Now try to fit a line which also goes through the origin



- Suppose we start with line L_1 then L_2 to L_5 (Rotate the line, find the best fit line and also it should go through the origin)
- Suppose line L_4 fits best.

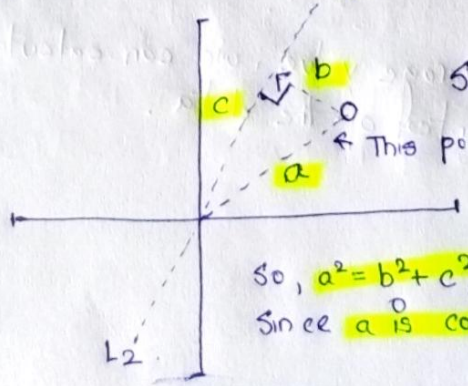
To quantify how good this line fits the data, PCA projects the data onto it. Suppose for example



or it can try to find the line that maximizes the distance from projected points to the origin. (maximize the distance)

It can either measure the distance from the data to the line and try to find the line that minimize those points. (minimize the distance)

Suppose let take a point to understand the minimize and maximize



So, a is constant (from the origin, never change).

This point is fixed from origin no matter what we change.

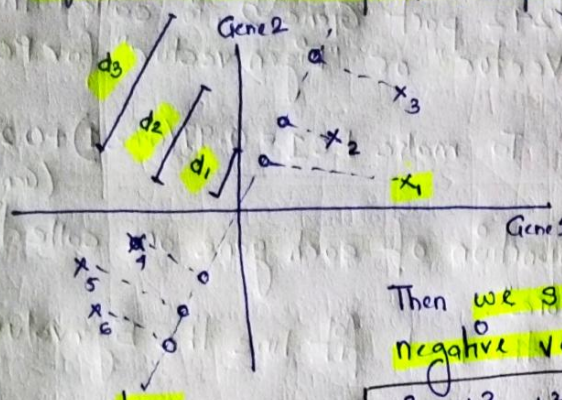
So when we project the point on the line we get Right angle triangle everytime.

So, $a^2 = b^2 + c^2$, pythogorous theorem where b and c are inversely related

Since a is constant, so if c get bigger, b get smaller (minimize)

& if c get smaller then b get bigger (maximize)

→ PCA find the best fitting line by maximizing the sum of the squared distance from the projected points to the origin.



PCA measure the distance from origin.
Suppose for x_1, distance from origin (from the best fit line) is d_1
So for all points $d_1, d_2, d_3, d_4, d_5, d_6$
Then we square the d_i so tve value don't cancel out negative values.

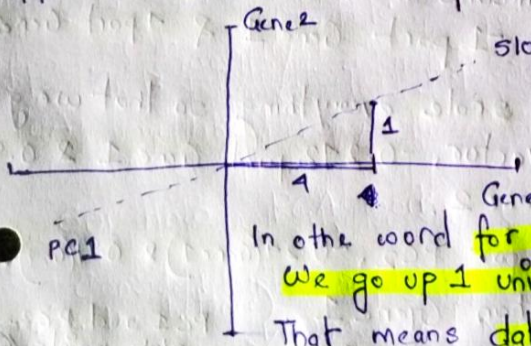
$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{Sum of square distances} = SS(\text{distances})$$

So we change the line L_2 to L_3 to L_4 and we repeat until we end up with

- the line with the largest (maximize) sum of squared distances between the projected point and the origin.

So we have a line (best line) which has the largest SS (distances)
And this line will be called as Principal Component 1 (PC1).

Suppose, PC1 has a slope of 0.25 means. $0.25 = \frac{1}{4}$

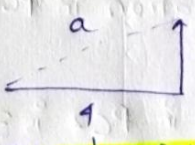


- PC1

In other word for every 4 units we go out along Gene 1 axis
We go up 1 unit along Gene 2 axis.
That means data are mostly spread out along Gene 1 axis and only a little bit spread out on Gene 2 axis.

So to make PC1 = Mix 4 part of Gene 1 with 1 part of Gene 2
In other word, Gene 1 is more important than Gene 2 when it comes to describing how data are spread out.
↳ This is also known as linear combination. So we can say PC1 is a linear combination of variables.

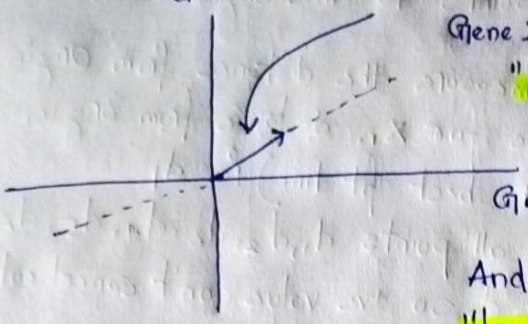
$a^2 = b^2 + c^2$, $a^2 = 4^2 + 1^2$, $a = 4.12$.



So when we do PCA with SVD, PC1 is scaled so length = 1 (a)
So divide each side by 4.12 ⇒ $\frac{4.12}{4.12}, \frac{1}{4.12}, \frac{4}{4.12}$ (just scaling)

So in scale version, PC1 = 0.971 (Gene 1) + 0.242 (Gene 2) 1, 0.242, 0.947.
but ratio is still same, 4 times as much as Gene 1 to Gene 2.
If length = 1, then this is a Unit Vector also.

Terminology →



This 1 unit long vector, consisting of 0.97 parts Gene 1 and 0.242 part Gene 2, is called the "Singular Vector" or "Eigen vector" for PC1.

To make PC1 = 0.97(Gene 1) + 0.242 (Gene 2)

And the proportion of each gene are called "loading score".

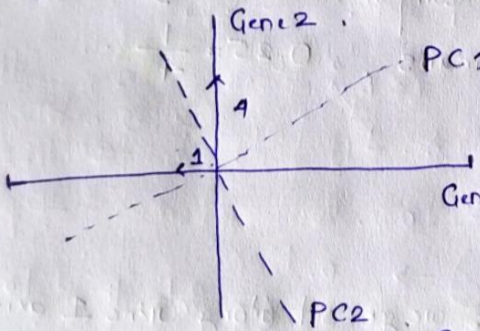
Also PCA calls the SS (distances) for the best fit line the Eigenvalue for PC1

$$\text{SS (distances for PC}_1\text{)} = \text{Eigenvalue for PC}_1.$$

$$\sqrt{\text{Eigen value for PC}_1} = \text{Singular Value for PC1}.$$

So lets work for PC2.

This is only a 2D graph, PC2 is simply the line through the origin that is perpendicular to PC1, without any further optimization that has done.



slope was 0.25 (1 and 4).

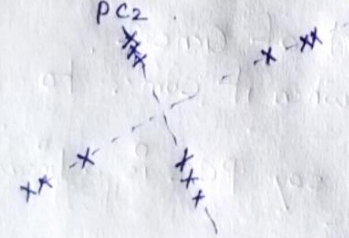
This means than PC2 is -1 part Gene 1 & 4 part Gene 2

If we scale everything so that we get a unit vector, -0.242 part Gene 1 & 0.97 part Gene 2.

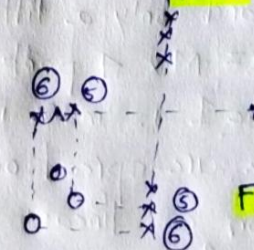
So loading score for PC2, -0.24(Gene 1) & 0.97(Gene 2)

So in PC2, Gene 2 is 4 times important than Gene 1.

So we have PC1 and PC2,



Rotate through → so that PC1 is horizontal



PC1 and PC2 plot points (6 point)

FINAL PCA plot

We can also check the variation in PC1 and PC2

$$\frac{\text{SS (distances for PC1)}}{n-1} = \text{Variation for PC1} = 15.$$

So variation in PC1 is 15.

$$\frac{\text{SS (distance for PC2)}}{n-1} = \text{Variation for PC2} = 3.$$

So variation in PC2 is 3.

So total variation is 15+3 = 18.

That means PC1 accounts for 15/18 = 0.83 = 83% of total variation around PC's.

So PC 2 accounts for 3/18 = 0.17 = 17% of total variation around PC3.

Suppose we have 3 variables (3 gene).

	M1	M2	M3	....	M6
Gene 1					
Gene 2					
Gene 3					

so we build $PC_1 = 0.62(\text{Gene 1}) + 0.15(\text{Gene 2}) + 0.77(\text{Gene 3})$

In this case Gene 3 is most important in PC_1.

Then find PC2, next best fitting line given that it goes through origin and perpendicular to PC_1.

for $PC_2 = 0.77(\text{Gene 1}) + 0.62(\text{Gene 2}) + 0.15(\text{Gene 3})$, Gene 1 is most important

In this way find PC_3. So if we have more variable find more & more principal components by adding perpendicular lines.

In practise, number of PCs is either number of variables or number of samples whichever is smaller.

Once we have all principal component figured out, we can use the eigen values (i.e. SS(distance)) to determine proportion of variation each PC account for.

Suppose variation in $PC_1 = 80\%$ $PC_2 = 14\%$ $PC_3 = 6\%$, then PC_1 and PC_2 accounts 94% variation in data. so choose only two PC_1 and PC_2.